Bayesian Analysis of Variability and Uncertainty of Arsenic Concentrations in U.S. Public Water Supplies

John R. Lockwood[1],
Mark J. Schervish[1],
Patrick L. Gurian[2]
and Mitchell J. Small[3]

The risk of skin and other possible cancers associated with arsenic in drinking water has made this problem a top priority for research and regulation for the U.S. EPA, as part of implementation of the Safe Drinking Water Act amendments of 1986 and 1996. To assess the costs, benefits and residual risks of alternative maximum contaminant levels (MCL's) for arsenic, it is important to characterize the current national distribution of arsenic concentrations in the U.S. water supply. This paper describes a Bayesian methodology for estimating this distribution and its dependence on covariates, including the source region, type (surface vs. ground water) and size of the source. The uncertainty of the fitted distribution is also described, thereby depicting the uncertainty in the proportion of utilities with concentrations above a given MCL. This paper describes the first stage of this assessment, based on a sample of concentrations from source water drawn by utilities. Subsequent analyses will incorporate the distribution and effectiveness of current treatment practices for reducing arsenic, and include available data sets of finished water quality to estimate the arsenic concentration distribution in water supplied to consumers.

Using arsenic concentration data for source (raw) water reported by 441 utilities from the National Arsenic Occurrence Survey (NAOS) (Frey and Edwards, 1997), we fit a Bayesian model to describe arsenic concentrations based on source characteristics. The model allows for both the formation of a national estimate of arsenic occurrence and the quantification of the uncertainty associated with this estimate. The specification of the model is

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

where

- Y_{ij} is the natural logarithm of arsenic concentration in μg/L at j^{th} source in i^{th} region
- μ_i is a constant for i^{th} region, where i ranges over the seven geographical regions specified in NAOS
- x_{ij} is the natural logarithm of the population served by j^{th} source in i^{th} region (an indicator of the size and flow rate of the utility source)
- g_{ij} is 0 if j^{th} source in i^{th} region is a surface water source and 1 if it is a ground water source

[1]Department of Statistics, Carnegie Mellon University.
[2]Department of Engineering and Public Policy, Carnegie Mellon University.
[3]Departments of Engineering and Public Policy and Civil and Environmental Engineering, Carnegie Mellon University.

- ϵ_{ij} represents those sources of random variation present at the j^{th} source in i^{th} region but not captured by the covariates in the model.

Furthermore, we model the values μ_i as independent normal random variables with mean ψ and variance τ^2. The national distribution of arsenic in source water is thus modeled as a mixture of lognormals with the mean of the log-concentration equal to $\mu_i + \beta x_{ij} + \gamma g_{ij}$ and the standard deviation of the log-concentration equal to σ. The resulting distribution depends upon the number of utilities in each of the seven regions (i), their service populations x and the respective numbers drawing water from surface ($g_{ij} = 0$) vs. ground ($g_{ij} = 1$) water (for now, the sample is assumed to be representative of the national distribution, though the predicted distribution can be readily modified to reflect a different distribution of the covariates in the target population).

To characterize the uncertainty of the fitted national distribution, we use vague prior distributions for the parameters ψ, τ, β, γ, σ and employ the Markov Chain Monte Carlo methodology (Gilks et al., 1996) to compute and simulate realizations from the posterior distribution of the parameters. Posterior uncertainty distributions of all quantities of interest can be calculated from these realizations.

Table 1 lists the posterior means and posterior standard deviations for the fitted model parameters. The mean values indicate that

- arsenic concentrations are generally higher in the west than in the east (the posterior means of μ_4, μ_5, μ_6 and μ_7 are greater than the posterior means of μ_1, μ_2 and μ_3)
- arsenic concentrations tend to be higher in source waters of larger utilities (the posterior mean of β is positive)
- arsenic concentrations are higher in ground water than in surface water (the posterior mean of γ is positive, though there is significant uncertainty in this result since the posterior standard deviation of γ is greater than the posterior mean)

The uncertainty in the fitted national distribution is characterized by the standard deviations of the parameters shown in Table 1 and by the covariance of the parameters in the posterior joint distribution. Figures 1 and 2 illustrate this covariance for two of the parameter pairs: (β, ψ) and (β, γ), respectively. These covariances are of the type that commonly arise in parameter estimation; for example, the positive association between higher β (which results in higher predicted arsenic concentrations) and lower ψ (which corresponds to lower values of the μ_i and lower predicted arsenic concentrations) is necessary to maintain the match to the observed sample values.

The national distribution is synthesized by sampling the joint parameter space (i.e, the points in Figures 1 and 2 and the associated points for the other model parameters) to generate many possible distributions. For each, the cumulative distribution function (cdf) at a particular value of the arsenic concentration ($\exp(Y)$) is computed as the average of the predicted cdf's for each measurement in the original sample of 441, based on its model covariates (or, the covariates for each utility in the target population, if these differ from the sample). The multiple cdf's generated from the parameter space describe the uncertainty of the national variability distribution. The median of the uncertainty distribution is one

Table 1: Posterior means and standard deviations of parameters. The regions (subscripts) are 1=New England, 2=Mid-Atlantic, 3=Southeast, 4=Midwest Central, 5=South Central, 6=North Central, 7=West.

Parameter	Posterior Mean	Posterior Standard Deviation
μ_1	-3.18	0.67
μ_2	-3.51	0.62
μ_3	-3.66	0.63
μ_4	-1.78	0.59
μ_5	-1.89	0.62
μ_6	-1.10	0.67
μ_7	-1.47	0.64
σ^2	2.17	0.20
ψ	-2.30	0.76
τ^2	1.74	1.77
β	0.21	0.05
γ	0.14	0.19

choice for a single estimate of the national distribution. This median distribution is shown in Figure 3, along with corresponding 5th and 95th percentiles and the observed distribution of the original data set. The fitted distribution closely matches the observed distribution, including the result that 37% of the sample is at or below the arsenic detection limit of 0.5 μg/L. The full uncertainty distribution for the proportion of the national population below one particular value of the arsenic concentration (5 μg/L) is shown in Figure 4, where this proportion is indicated to range from about 0.79 - 0.87, with a median of 0.83. This characterizes the uncertainty in the proportion of utilities requiring treatment of their source water to meet an MCL of 5 μg/L.

Acknowledgment: This work is sponsored by the U.S. EPA Office of Ground Water and Drinking Water, Standards and Risk Management Division. The paper has not been subject to EPA peer review, and the views expressed are solely those of the authors.

References

Frey, M. M. and M. A. Edwards (1997). Survey arsenic occurrence. *Jour. AWWA*, **89**(3), 105-117.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, eds (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

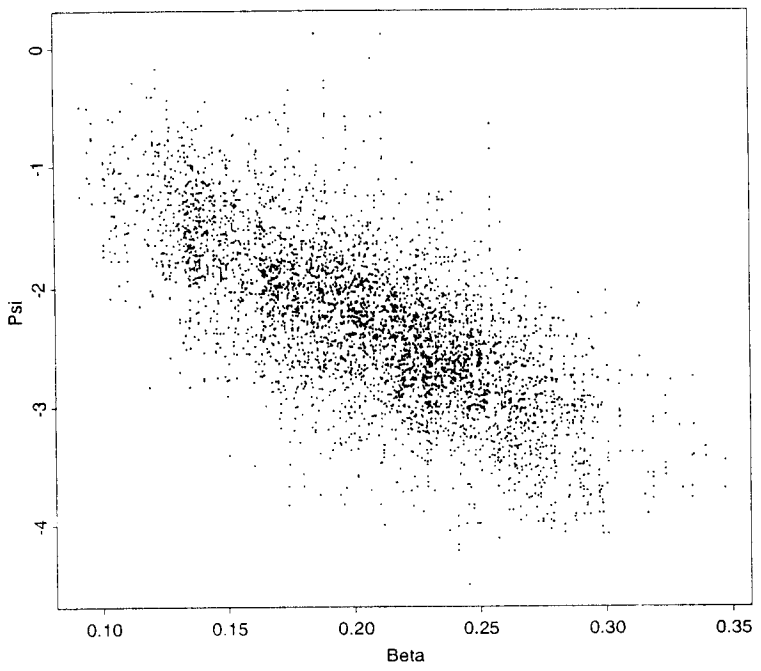

Figure 1: Scatterplot of ψ versus β from a sample of size 5000 from the joint posterior distribution

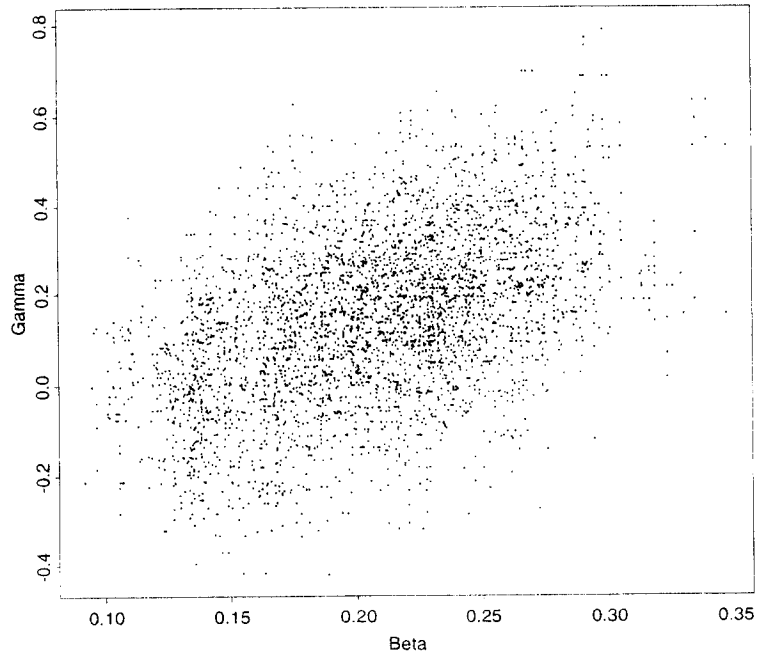

Figure 2: Scatterplot of γ versus β from a sample of size 5000 from the joint posterior distribution

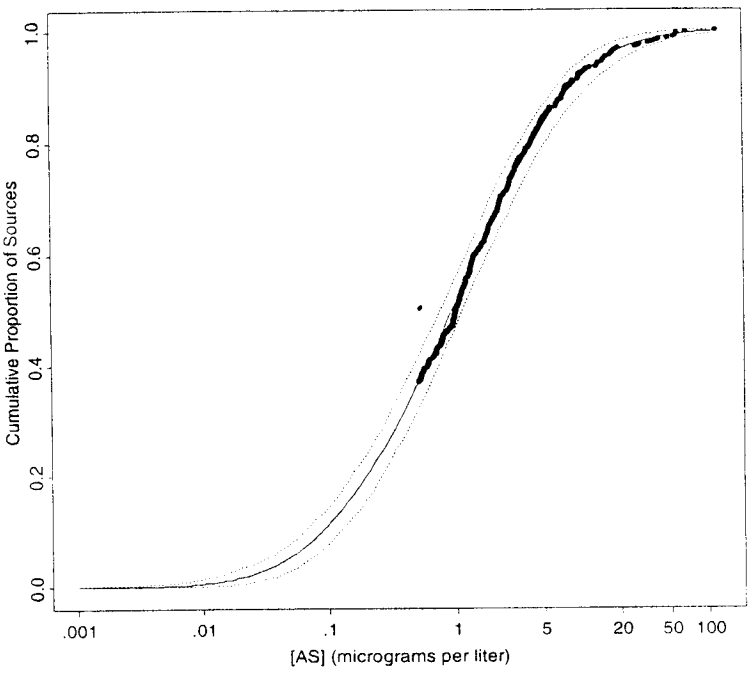

Figure 3: Posterior cumulative distribution function of national arsenic occurrence in source water with 90% credible bounds and uncensored NAOS data overlayed.

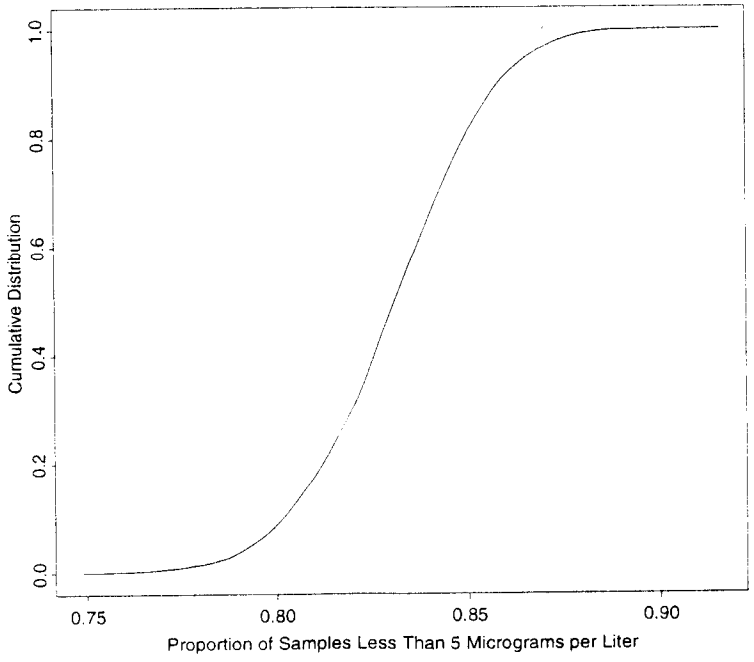

Figure 4: Posterior cumulative distribution function of the proportion of national arsenic occurrence less than 5 μg/L